VOLUME 64
NUMBER 11

WHOLE NO. 317
1950

Psychological Monographs:

General and Applied

Combining the *Applied Psychology Monographs* and the *Archives of Psychology* with the *Psychological Monographs*

HERBERT S. CONRAD, *Editor*

An Evaluation of Personality-Trait Ratings Obtained by Unstructured Assessment Interviews

By

ERNEST C. TUPES

Human Resources Research Center
Lackland Air Force Base
San Antonio, Texas

Based on a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the University of Michigan
Committee in charge: G. A. Satter, *Chairman;* E. S. Bordin, E. L. Kelly, W. C. Trow, and E. L. Walker

Accepted for publication, May 8, 1950

Price $1.00

Published by

THE AMERICAN PSYCHOLOGICAL ASSOCIATION
1515 MASSACHUSETTS AVE. N.W., WASHINGTON 5, D.C.

ACKNOWLEDGMENTS

THE author wishes to thank his entire committee, and especially Professors George A. Satter and E. Lowell Kelly, for their guidance and constructive criticism in the present study. Additional gratitude is expressed to Dr. Kelly for making available data collected by The Research Project on the Selection of Clinical Psychologists. A special debt of gratitude is owed to the author's wife, Wanda, who by her constant encouragement the past three years, made the present study possible.

ERNEST C. TUPES

TABLE OF CONTENTS

AN EVALUATION OF PERSONALITY-TRAIT RATINGS OBTAINED BY UNSTRUCTURED ASSESSMENT INTERVIEWS

I. PROBLEM, PROCEDURES, AND METHODS

THE TERM "Interview" as usually defined means almost any face-to-face contact between two or more individuals which involves the exchange of information. When the purpose of the interview is to obtain information about the person being interviewed, it is usually called a diagnostic or appraisal interview. It is with this type of interview that this investigation is concerned.

As pointed out by Fearing (6, 7), the appraisal interview is "probably one of the oldest human social techniques," and is also "the least studied, the most constantly used, and the most frequently challenged method of securing social data."

Studies of the reliability of judgments based on the interview (2, 7, 9, 10, 14, 15), although extremely divergent with respect to method of interview and the type of data selected for analysis demonstrate that these judgments will vary in reliability according to: (*a*) the person doing the interview; (*b*) the amount of objective data available to the interviewer before the interview; and (*c*) the degree of structuring or standardization of the interview itself.

The validity of judgments based on the interview technique has been less adequately studied than has the reliability of interview judgments. In 1931, Symonds (17) pointed out that interviewing "has not been subjected to experimental scrutiny or statistical validation." A survey of the literature by the present writer did not indicate that the situation has been changed since 1931. In none of the studies (1, 4, 5, 10, 13) was any estimate made of the contribution of the interview itself to the validity of the interviewer's judgments. In every case the validities given were for judgments based on the interview plus other material. As a group, the validity studies indicated that little evidence has been gathered to support the belief, prevalent among psychiatrists and psychologists as well as persons in general who deal with other people, that face-to-face contact in an interview situation is very necessary (and sometimes sufficient) for measuring personality traits and predicting behavior.

The present investigation is an attempt to determine the validity of personality-trait ratings based on two types of unstructured interview situations and, further, to determine the increase in validity of interview ratings over that of ratings made by the same individuals without benefit of an interview. The data analyzed in this study were gathered during the 1947 Michigan Assessment Program.[1] If in some instances the data appear incomplete or the results inconclusive it should be understood that the Assessment Program was not designed[2] to thoroughly investigate any particular technique but rather to determine the maximum validity of many techniques combined.

A. SUBJECTS

The sample investigated in this study consisted of 128 male college graduates who had been accepted by various universities for training in clinical psychology leading to the Ph.D. degree under the Veterans Administration training program. All were beginning graduate students accepted at the P-1 level—the lowest professional grade in the U. S. Civil Service system.

[1] The 1947 Michigan Assessment Program was conducted as a part of the Research Project on the Selection of Clinical Psychologists sponsored by the Veterans Administration under a contract with the University of Michigan. For a detailed report of the overall aims and research design of this project, the reader is referred to the project's Preliminary Report issued in December, 1948 (11).

[2] The writer, although associated with the project since September, 1946, had no part in the planning or experimental design of the 1947 Michigan Assessment Program and wishes to accept no credit therefor. Neither does he wish to be held responsible for certain justifiable criticisms of the design such as the lack of estimates of the reliability of interview ratings and the lack of independence between interview ratings and criterion ratings.

B. The Raters

The subjects were studied intensively over a period of one week by a staff of thirty clinicians. Two of these were psychiatrists. The majority of the others were professional clinical psychologists, some being advanced graduate students in clinical psychology. Ratings made by these clinicians comprise the data of this investigation.

C. The Rating Scale

The rating scale used in the Assessment Program comprises 42 variables divided into a Scale A (a group of 22 bi-polar so-called surface traits—supposedly the more manifest dimensions of personality), a Scale B (a group of 9 so-called source traits—supposedly the more underlying dimensions of personality), and a Scale C (a group of 11 so-called criterion variables, which in reality are various skills believed necessary for successful functioning as a clinical psychologist). In addition, ratings were made on *Liking* (Variable #0), defined simply as extent of personal like or dislike for the subject. In rating students on Scales A and B the staff was instructed to rate the person as he was at the time of assessment. In rating on Scale C the student was to be rated as he would be five years in the future (i.e., when he would have one year of professional experience past the Ph.D. degree). Definitions of Scales B and C variables are given in the Appendix.

All three rating scales were the product of joint thinking on the part of the project planning committee and were the outgrowth of trial scales used in preliminary assessments the year before. Scale A was based in part on the findings of Cattell (3) in his factor analyses of personality ratings.

In using the rating scales, raters were instructed to use as a frame of reference "all first year clinical psychology graduate students at universities accredited by the American Psychological Association to offer training in clinical psychology." Raters were asked to make their ratings conform roughly to a normal distribution with suggested frequencies designated as 3, 7, 15, 25, 25, 7, and 3 per cent, respectively, for each point from 1 to 8.

D. The Interviewers

A staff team of three members studied each student intensively. This team consisted of an Initial Interviewer, an Intensive Interviewer, and either a Test Integrator or a Projective Integrator. Each staff team studied four students from each class of 24 students. Sixteen staff member functioned as Initial and Intensive Interviewers interchangeably so that for each student team of four students, one staff member would act as Intensive Interviewer for two students and as Initial Interviewer for the other two students. The group of 128 students studied in this investigation thus received two sets of interview ratings—one set made after the Initial interview (hereinafter designated as *Init* ratings) and one set made after the Intensive interview (hereinafter called *Intens* ratings). Since the 16 interviewers acted in both the Initial and Intensive interviewer capacity for the 128 students (although no staff member functioned as both Intensive and as Initial Interviewer for the same student), the problem of interviewer differences does not arise when comparisons are made between ratings based on the Intensive interview situation and ratings based on the Initial interview situation. Each of the interviewers interviewed from two to twelve students in

the role of Initial Interviewer and a similar number in the role of Intensive Interviewer.

All 16 interviewers (2 psychiatrists and 14 professional clinical psychologists) had had considerable interviewing experience before assessment. Without doubt they may be regarded as being more skilled at uncovering personality dynamics by use of the unstructured interview than the general population of interviewers. In fact all could undoubtedly be termed expert interviewers.

E. The Initial Interview Situation

The Initial Interview was an unstructured type of interview lasting approximately one hour. The material covered in this interview was purposely kept at a fairly superficial level so that little anxiety would be aroused in the subjects. The Initial Interviewer had available to him before the interview only the information about the subject usually found in a credentials file: i.e., application blank (Civil Service Form 57), letters of recommendation, and records of past performance in the form of college transcripts. The Initial interviewer is probably comparable to the usual college interview for the purpose of determining admission elegibility, and is similar to the typical interview (although somewhat longer in duration) given applicants for employment in industry.

F. The Intensive Interview Situation

The Intensive interview was also unstructured. It lasted approximately two hours. The Intensive Interviewer had available before the interview the wealth of objective, projective, and autobiographical material described in detail below under *Design of Assessment* (Section I, G, 1). The Intensive interview was deep and probing in nature: the interviewer attempted to uncover underlying personality dynamics in order to confirm hypotheses suggested by his review of the data available on the student, or to fill gaps which persisted in spite of all the data available. The Intensive interview is probably comparable to the interview used for diagnostic purposes with a neuropsychiatric patient by a psychiatrist who has available to him the social history, letters from friends and relatives, and psychological test reports.

G. The Design of Assessment

Lack of space prevents the presenting of the experimental design of the 1947 Michigan Assessment Program in detail.[3] However, some information concerning the assessment is desirable so that the role of the interviews may be seen in proper perspective. Therefore, a typical student ("X") will be followed through the assessment program. In addition, the various assessment ratings will be described in some detail.

1. *The students' assessment schedule.* Students came to the assessment center in groups of 20 to 24 for periods of seven days. Soon after his arrival at assessment headquarters, "X" and his classmates were given an orientation by the Project Director, at which time the aims of assessment were reviewed and the students assured that personal material would be kept in strictest professional confidence. "X" was also told that since certain staff members would be rating him on the basis of certain specified materials he would have absolutely no contact with those staff members until after those particular ratings had been made. The students were then assigned to teams of four.

During his first several days of assessment, "X" took a battery of paper-and-pencil objective tests (Miller Analogies Test, Chicago Test of Primary Mental Abilities [Single Booklet Edition], Allport-Vernon Study of Values, Strong Vocational Interest Blank, Minnesota Multi-

[3] For a complete description of the assessment see the Preliminary Report (11) referred to earlier.

phasic Personality Inventory, and the Guilford-Martin Battery—Inventory of Factors STDCR, Inventory of Factors GAMIN, and Personnel Inventory I).[4] Scores on these instruments were made available to the assessment staff. The assessee was given a projective test battery consisting of the Rorschach, the Bender-Gestalt, a sentence completion test, and ten cards of the Thematic Apperception Test. He filled out a biographical inventory (consisting of 131 items), a psychological experience record (for aid in evaluating objective and projective tests), and wrote from an outline a detailed autobiography (average length about 20 pages). He was interviewed by an Initial Interviewer, and later interviewed again by an Intensive Interviewer.

On the fifth day of assessment, "X" was put through a period of situation or role-playing tests designed to bring out salient features of the student's interactions with others.[5] He was observed and rated at this stage by his own staff team (the two interviewers and the Projective or Test Integrator), and was also observed and rated by another staff team who had never seen him before and who had no information available about him. That evening, all students and staff relaxed with a social gathering.

After breakfast on the last morning, "X" was asked to fill out a sociometric questionnaire (typical questions: "Who would you most like to go on a camping trip with?" and "Which student would you prefer to be supervised by in your work?") concerning his reactions to his classmates. In addition, he was asked to rate himself and his three teammates on Scales A, B, and C (using a slightly modified form of the staff rating form). Finally he was asked to prepare concise but frank character sketches about his teammates.

When these tasks were completed, "X" had a final appointment with his intensive interviewer. Although he hoped to obtain from that interview the staff's opinion of him, he went away satisfied with a small amount of relatively innocuous data, such as test scores.

Finally, "X" and his classmates met for a group-therapeutic session conducted by a visiting psychologist (who had taken no other part in the assessment program) and a farewell talk by the Project Director, who thanked the group for their cooperation and repeated the assurance that assessment findings would be kept in strictest confidence.

2. *Ratings made during assessment.* The preceding section indicates the wide variety of psychological techniques used in assessment. Ideally, the program would have been designed so that each technique could be evaluated separately, with staff members rating independently each student on each technique. Unfortunately, the expense of such a design (in terms of staff and student time as well as money) was prohibitive. As a compromise it was decided to make ratings on various combinations of psychological data and to design the assessment program so that even though estimates of the validity of ratings based on all techniques might not be obtained directly, it would be possible to estimate the increase in validity of ratings obtained with the addition of different psychological data.[6] Wherever practical, in-

[4] In addition to these objective tests, the students took the Kuder Preference Record, the ACE General Culture Test, and the Strong Vocational Interest Blank twice more—once with instructions to fill it out as it would be answered by "women in general" and finally as it would be answered by "men in general." Scores on these tests were not made available to the assessment staff but were filed away for later independent analyses.

[5] The situation tests were patterned after the psychodramatic (role-playing) technique of Moreno. For a complete description and thorough analysis of situation tests as used in the 1947 Michigan Assessment Program, the reader is referred to an article by Dr. William Soskin (17). One typical situation test involved two subjects, one of whom was informed in advance that he was to play the role of a small-town high school principal who had called one of his male instructors in for a conference regarding the instructor's rumored sexual misconduct. The other subject was informed that he was to play the role of a high school instructor who had been told the principal wished to talk to him but was not informed what the conference was about. The ensuing five or ten minutes of role-playing consistently called forth aspects of the students' personalities that surprised and enlightened even the participants.

[6] Immediately after each set of ratings was made, the ratings were filed away in the project office and were not again made available to any staff member until the Final Pooling Conference. Similarly, notes made by any rater were not available to any other rater nor were the students discussed with other raters until the Preliminary Conference. Thus, the Intensive Interviewer, for example, when making ratings after the Intensive interview had available only

dependent ratings on various separate psychological techniques were obtained as well. In general, the psychological data were utilized in order of increasing cost; credentials material costing only a postage stamp was presented to the raters first; objective tests costing only slightly more were presented second, etc.

The following ratings were made during assessment:

a. Ratings based on each of the four projective techniques made by the staff member administering the particular technique.

b. Ratings based on the four projective protocols plus interpretations of the protocols. These ratings were made by the Projective Integrators on half the students.

c. Ratings made on the other half of the students by the Test Integrators. These ratings were based on all projective material and in addition on all credentials material, all autobiographical material and all objective test data.

d. Ratings made by the Interviewers.

(1) *PreInit* ratings. These ratings were made by the staff member serving as Initial Interviewer prior to the interview —i.e., before the raters had seen the subjects. These ratings were based entirely on the material contained in the Credentials folders.

(2) *Init* ratings. These ratings were made by the Initial Interviewers immediately after the Initial interviews. Thus, they were based on Credentials materials plus the information obtained during the Initial interview.

(3) *PreIntens* ratings. These ratings were the fourth set of ratings made by the Intensive Interviewers before seeing the subjects. The Intensive Interviewers had previously rated on Credentials material alone; then on Credentials plus Objective test data; then on Credentials plus Objectives plus Autobiographical material. The *PreIntens* ratings were based on Credentials material, Objective test data, Autobiographical material, and Projective protocol material.

(4) *Intens* ratings. These ratings were made by the Intensive Interviewers immediately after the Intensive interview. Thus, they were based on the data described under *PreIntens* ratings above plus the information obtained during the Initial interview.

e. Ratings decided upon at the Preliminary Pooling Conference which occurred after the Intensive Interview. The team of three staff members (the two interviewers and the Projective or Test Integrator) assigned to the student participated in this conference. All the data referred to above were available for this conference as were the notes taken by any of the three staff members prior to the conference. None of the ratings made previously on the student either by the three staff members or by any other staff members were available for the conference.

f. Ratings based on the Situation tests. There were of two types: The three sets of individual Contaminated Situation ratings made by the student's own staff team, and three sets made by another staff team who knew nothing about the student and thus were forced to base their ratings solely upon impressions gained from the Situation tests. In addition, this latter staff team held a Situation Pooling Conference at which they arrived at a set of Uncontaminated Pooled Situation ratings.

---

the written test material and the notes he himself had made during the interview. He had no knowledge of the ratings or the opinions of any other person.

g. The Final Individual ratings. On the last morning of assessment, each member of the student's staff team reviewed all material available on the student (including anything significant he might have gleaned during the social gathering following the Situation tests), and made a set of Final Individual ratings.

h. Ratings by the students. On the last morning of assessment each student rated himself and his three teammates.

i. The *FinP* ratings. During the last afternoon of assessment the student's staff team met for a Final Pooling Conference at which was made available not only all psychological material gathered during assessment, but all ratings that had been made on or by the student. The *FinP* ratings arrived at during this conference thus represent the combined judgment of three clinicians who had intensively studied the subject for seven days and who had available a wealth of psychological data.

H. Assessment Data Used in This Investigation

The *PreInit*, *Init*, *PreIntens*,[7] and *Intens* ratings described in the preceding section comprise the ratings whose validities are estimated in this investigation. These ratings were validated against the *FinP* ratings. The estimated reliabilities of the *FinP* ratings are presented in Table 1 (in Section II, *Results*). Unfortunately, the assessment design did not allow for any estimate of the reliabilities of the predictor ratings.

[7] The other three sets of ratings made by the Intensive Interviewers before the interview (see Section G,2,*d*,3 above) are being analyzed elsewhere. Their analyses are not included here because these ratings did not seem especially relevant to the main purposes of this study—the investigation of the validity of ratings based on interviews and increase in validity which can be attributed to the interviews.

The present paper is concerned only with the validity of ratings on Scales B and C. Scale A was not rated by either of the interviewers prior to the interview, and to present the validities of the interview ratings for Scale A would add nothing to the conclusions drawn from the investigation.

I. The Criterion Measures

Since criterion measures are of paramount importance in any validity investigation (and this is especially true in the area of personality measurement where good criteria are rare indeed), the *FinP* ratings should be examined in detail to determine whether they are acceptable criteria of the personality variables rated.

The estimated coefficients of reliability of the *FinP* ratings (Column 5 of Table 1) on the Scale B variables ranged from .73 to .90 (median .87); those on Scale C from .88 to .93 (median .90). These reliabilities, while not as high as is desirable for individual evaluations, are satisfactory for group measurements. Certainly, they are of such magnitude as to permit evaluation of predictive devices; i.e., failure of any predictor to correlate highly with the *FinP* ratings can only be interpreted as lack of validity of the predictor.

In addition, as noted above, the *FinP* ratings represent the combined and pooled judgments of three staff members, selected on the basis of professional competence in the field of clinical psychology or psychiatry, who had intensively studied each subject for a period of one week, making use of a wide variety of psychological techniques and materials.

Considering these facts, there would seem to be little doubt that the *FinP* ratings of the personality variables making up Scale B are about as valid cri-

terion measures of these variables as are attainable from trained clincians with present techniques.

The Scale C variables are a different matter. Scale C variables are not personality traits but rather predictions regarding future performance (five years later) in various aspects of the job of Clinical Psychology; Variable C_{42} was an appraisal of the overall suitability of the student for the field of clinical psychology. It is not believed, therefore, that Scale C *FinP* ratings are acceptable as criterion measures, even though they may represent "better guesses" than ratings made earlier in the assessment program. For this reason comparisons between ratings on Scale C variables based on the interview situations and *FinP* ratings will be considered merely as agreements between two sets of ratings.

One factor which must be considered, if the *FinP* ratings are to be used as criterion measures, is the factor of contamination of the *FinP* ratings. Of the three persons comprising the staff team who made the *FinP* ratings, two were the interviewers whose ratings are here being investigated. If the role of either Intensive Interviewer or Initial Interviewer was regarded consistently by the other two team members as being a "better" role for judging personality characteristics, then that interviewer would have a dominance in the pooling conference which would tend to influence the other two team members. The effect of this influence would be to consistently bring the *FinP* ratings closer to the ratings made by the interviewer than would be warranted by the actual merit of those ratings. Dominance by individuals apart from their roles may be disregarded since all interviewers acted in the dual capacity of Intensive and Initial Interviewers on the same staff team. Staff members were rotated from week to week so that for each student class the staff teams were made up of different combinations of staff members; this also tended to reduce the effect of dominance by any particular individuals.

Detailed evidence on the effect of such role dominance on the estimates of validity of the interviewer's ratings will be presented elsewhere (18). The evidence, while not conclusive, does indicate that role dominance was present; that the role of Intensive Interviewer was believed by the other staff members to be especially advantageous for judging personality variables; but that the effect of such dominance on the *FinP* ratings was slight. The possibility that Intensive Interviewer validity coefficients are spurious will be considered when the results of this investigation are discussed in Section III.

J. Method of Determining Validity of Ratings Based on Interviews

To determine the validity of ratings on the various traits made by the interviewers either before or after the interview, the rating given each subject by the interviewer on a particular variable was plotted against the criterion rating given that subject on the same variable. Pearson product-moment correlation coefficients were then computed from the scatter plots, using the conventional raw-score formula. The various sets of correlation coefficients obtained are given in Table 1 appearing in the following chapter.

To determine the contribution of the interview itself to the validity of the ratings made after the interview—that is, the contribution of the interview apart from the material available to the interviewer—correlation coefficients between

ratings made by the interviewer *before* the interview and the criterion ratings were squared and these squares subtracted from the squares of the coefficients of correlation between ratings made *after* the interview and the criterion ratings. The resulting figures may be thought of as the percentage of variance in the *FinP* ratings which can be accounted for on the basis of the interview. Since these figures are a function of how validly the variable was rated before the interview (variables more validly rated before the interview being more limited in the amount the correlation coefficient could increase after the interview) an additional correlation was made which will be discussed in Section II.

II. RESULTS

THE MAIN findings of this investigation are presented in Table 1. In discussing the findings, those regarding the Initial interview will be considered first; those concerned with the Intensive interview second; and those concerned with agreement between Initial and Intensive interviews last.

A. The Initial Interview

The mean ratings[1] assigned by the Initial Interviewers after the interview ranged from 4.32 to 5.61 (median mean rating 4.80). The criterion means ranged from 3.68 to 5.56 (median 4.42). The standard deviations of the *Init* ratings range from 1.08 to 1.61 (median 1.41). The standard deviations of the criterion ratings ranged from 1.13 to 1.56 (median 1.32).

1. *Validity of ratings made before the Initial interview.* In Column 1 of Table 1 are listed the correlations between the ratings, based on Credentials material only, made by the Initial Interviewers before they had seen the subjects (*PreInit* ratings) and the *FinP* ratings arrived at by the staff team after their Final Pooling Conference. These correlations are small (only eleven are significantly different from zero at the .01 level) and account for only a small amount of the criterion variance. As might be expected from the nature of the Credentials material (job histories, college transcripts, and the like), the variables most validly rated are those (B31, C32, C36, and C39) concerned with future intellectual performance.

2. *Validity of ratings made after the Initial interview.* The validities of the ratings made by the Initial Interviewers after the Initial interview are given in Column 2 of Table 1, and the estimated validities of these ratings, if the criterion ratings were perfectly reliable, are listed in Column 6.

It is evident that the Initial Interviewer after the interview was, on all variables, able to make ratings which correlated with the *FinP* ratings more closely than would have occurred had the *Init* ratings been made solely on a chance basis. All of the *Init-FinP* r's are larger than 23.

When the *Init-FinP* correlations are corrected for unreliability in the *FinP* ratings,[2] the increase in correlations is negligible. The median r for Scale B is raised from .42 to .47, and that for Scale C from .46 to .49. The rank order correlation (*rho*) between the corrected and uncorrected r's is .99, indicating that lack of reliability in the *FinP* ratings had little effect on the relative validity with which the different variables were rated by the Initial Interviewer. The *rho*'s between the *Init-FinP* correlations and the reliabilities of the *FinP* ratings are .17

[1] Only summary data concerning means and standard deviations of interview and criterion ratings are presented here. Complete tables of means and standard deviations will be presented in a forthcoming article (19). In the meantime, copies of these tables may be obtained from the author.

[2] The *FinP* reliabilities are given in Column 5 of Table 1. These reliabilities were estimated in this manner:

The reliabilities of the Final Individual ratings (made just before the Final Pooling Conferences) were first estimated by computing r_{ab}, r_{ac}, and r_{bc}, where a, b, and c are the sets of Final Individual ratings made by the three team members, respectively. The median correlation for each variable was selected as the best indication of the reliability of a single rater. The coefficients of reliability for the *FinP* ratings were estimated by the Spearman-Brown prophecy formula applied to these median correlations.

TABLE 1

SUMMARY OF RESULTS*

Variable	*PreInit-FinP* r	*Init-FinP* r	*PreIntens-FinP* r	*Intens-FinP* r	*FinP* Reliability	*Init-FinP* $r_{x\infty}$	*Intens-FinP* $r_{x\infty}$	*Init Net Gain* Expressed as Per Cent	*Intens Net Gain* Expressed as Per Cent	*Init-Intens* Agreement
(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
B 23. Social Adjustment	.02	.38	.50	.61	.90	.40	.64	16	18	.23
B 24. Appropriateness of Emotional Expression	−.07	.29	.49	.62	.85	.32	.67	10	23	.07
B 25. Intensity of Inner Emotional Tension	−.09	.50	.21	.55	.84	.54	.60	30	32	.41
B 26. Sexual Adjustment	−.01	.51	.49	.61	.88	.54	.65	30	20	.36
B 27. Motivation (Status)	.28	.42	.45	.54	.73	.49	.64	15	17	.16
B 28. Motivation (Science)	.27	.28	.52	.63	.88	.30	.67	01	21	.19
B 29. Insight into Others	.22	.32	.36	.50	.86	.34	.54	07	16	.14
B 30. Insight into Self	.21	.44	.48	.63	.87	.47	.68	18	27	.14
B 31. Quality of Intellectual Accomplishment	.40	.56	.70	.77	.90	.59	.81	21	24	.50
Median, Scale B	.21	.42	.49	.61	.87	.47	.65	16	21	.19
C 32. Academic Performance	.54	.65	.76	.78	.91	.68	.82	21	09	.45
C 33. Clinical Diagnosis	.32	.43	.59	.69	.90	.45	.73	10	24	.26
C 34. Individual Psychotherapy	.27	.50	.48	.63	.90	.53	.66	21	25	.24
C 35. Group Psychotherapy	.20	.40	.45	.65	.90	.42	.67	14	39	.31
C 36. Research	.47	.63	.68	.72	.90	.66	.76	26	14	.48
C 37. Administration	.25	.38	.49	.63	.88	.40	.67	10	25	.21
C 38. Supervision	.28	.50	.46	.66	.91	.53	.69	21	33	.29
C 39. Teaching	.39	.49	.68	.74	.90	.52	.78	12	20	.34
C 40. Professional Interpersonal Relations	.16	.37	.43	.62	.91	.39	.65	13	27	.10
C 41. Integrity	.10	.46	.60	.70	.89	.49	.74	23	25	.35
C 42. Overall Suitability	.29	.46	.57	.70	.93	.48	.73	15	28	.37
Median, Scale C	.28	.46	.57	.69	.90	.49	.73	15	25	.31

* $N = 128$. To be significant at the .01 level, r must be .23 or larger; to be significant at the .05 level, r must be .18.

for Scale B and .28 for Scale C. Apparently differences in reliability of the *FinP* ratings are only slightly related to differences in correlations between *Init* ratings and *FinP* ratings.

In order to test the significances of differences between the *Init-FinP* r's for all 20 variables, the r's were first converted into Fisher's *z-functions* by use of a table given by Lindquist (12, p. 212), and the significance of the difference between each pair of z's was estimated. All possible combinations of the 20 variables taken two at a time amount to 189 pairs of differences. Thirty-eight of these differences are significant ($P = .05$ or less). By chance one would expect to find only five differences out of each hundred (or nine out of the 189) reaching the .05 level of significance. The conclusion seems justified that differences do exist in the validities with which the different variables were rated by the *Init* interviewer.[3]

3. *The contribution of the interview to the validity of ratings made after the Initial interview.* Comparison of the *PreInit-FinP* correlations (based on ratings before the interview) with the *Init-FinP* correlations (based on ratings after the interview) provides an opportunity to estimate the incremental contribution of the Initial interview to the validity of the post-interview ratings. If the *PreInit-FinP* r's (Column 1 of Table 1) and the *Init-FinP* r's (Column 2 of Table 1) are squared, the differences between these r^2's provide estimates of the amount of *FinP* rating variance accounted for by the *Init* ratings which can be attributed to the Initial interview. However, the *FinP* ratings differed in reliability and it would seem that these "gain in variance accounted for" figures might be more representative of the contribution of the interview were they corrected in some manner for *FinP* reliability differences. Therefore, an *Init Net Gain* figure was computed by subtracting the *PreInit-FinP* r^2 to obtain a "possible gain in variance accounted for" figure by which the "gain in variance accounted for" figure was divided. It is these relative *Init Net Gains* which are listed in Column 8 of Table 1.[4]

For illustrative purposes, an example of the computation of the *Init Net Gain* figure for Variable B23 follows: From Table 1, it may be seen that for Variable B23, the *PreInit-FinP* r is .02 and the *Init-FinP* r is .38. The difference between the squares of these r's is .1440, which can be interpreted as the amount of the variance of the *FinP* ratings on B23 accounted for by the *Init* ratings which can be attributed to the Initial interview. The reliability of the *FinP* ratings on B23 is .90. Subtracting from

[3] This conclusion is rendered even more tenable by the fact that some correlation might be expected to be present between the z's (since the r's on which the z's are based were all computed on the same sample of cases and hence are not independent), but the amount of this correlation is unknown and thus was ignored in estimating the standard errors of the differences. The effect of ignoring such correlation is to overestimate the standard errors of the differences and thus to underestimate the significance of the differences.

[4] Partial correlation coefficients could have been used to determine the net contribution of the interview to the validity of the post-interview ratings, but when both pre-interview and post-interview validity coefficients are large, a greater difference between the first order r's is necessary to obtain the partial r which would be obtained by a smaller difference between first order r's when their magnitude is smaller. Partial correlation coefficients thus would cause an underestimation of the relative contribution of the interview when the pre-interview ratings correlated higher with the criterion ratings. The *net gains* are not influenced by the magnitudes of the r's, and in addition, by taking account of the reliability of the *FinP* ratings, do not penalize the interviewer's ratings on those variables where, because of lower *FinP* reliabilities, the maximum *Intens-FinP* correlation possible is actually less.

this the square of the *PreInit-FinP r* (.0004), the "possible gain in variance accounted for" figure .8996 is obtained. Dividing .1440 by .8996 results in the *Init Net Gain* figure .16, which is given in Column 8 of Table 1.

Examination of Column 8 indicates that there are appreciable differences in the *Init Net Gains* between the different variables, but interpretation of these differences is difficult. The two variables with highest *net gains* are B25 and B26, which were among the five variables (18) on which the Final Individual ratings of the Initial Interviewer had the highest (of the three team members) correlations with the *FinP* ratings, so role dominance may be a contributing factor. A common core underlying the variables (B24, B28, B29, C33, and C37) which gained relatively least from the Initial interview becomes apparent when it is considered that these variables all have fairly high loadings on the same factor[5]—one which was tentatively defined as *social intelligence*, or the *ability to effectively use intelligence in interpersonal relations*. Apparently the credentials folder furnished some clues to general intelligence (college transcripts, work history, etc.) which allowed the raters to rate these variables with some success before the interview. The interviewer was able to obtain little information during the Initial interview which added to the estimate of intelligence already made, and little other information which was of help in estimating how effectively the subjects were able to use their intelligence in social situations.

[5] This factor was obtained in an unpublished factor analysis of the intercorrelations between *FinP* ratings of Scales B and C carried out by the Research Project on the Selection of Clinical Psychologists.

B. The Intensive Interview

The mean ratings assigned by the Intensive Interviewers after the interview ranged from 3.89 to 5.54 (median mean rating 4.65). The standard deviations of the *Intens* ratings varied from 1.14 to 1.63 (median standard deviation 1.38).

1. *Validity of ratings made before the Intensive interview.* The correlations between the *PreIntens* ratings (the ratings, made by the Intensive Interviewers just before the interview, based on Credentials, Objective test data, Autobiographical material, and Projective protocols), and the *FinP* ratings appear in Column 3 of Table 1. With one exception (Variable B25) the correlations are significant at or beyond the .01 level of significance. As was found true for the *PreInit* ratings, the variables most validly rated by the Intensive Interviewer before the interview are B31, C32, C36, and C39—all variables which are concerned with intellectual performance.

2. *Validity of ratings made after the Intensive interview.* In Column 4 of Table 1 are listed the validities of ratings made by the Intensive Interviewer after the Intensive interview. In Column 7 appear the validities which would have resulted had the *FinP* ratings been perfectly reliable.

All of the *Intens-FinP* correlations attain significance beyond the .01 level.

It should be noted that correcting the *Intens-FinP r*'s for unreliability of the *FinP* ratings raised the median *r* from .61 to .67 for Scale B, and from .68 to .71 for Scale C. Even had the *FinP* ratings been perfectly reliable the *Intens* ratings would have done only a slightly better job of predicting the *FinP* ratings. The rank difference correlation coefficient, *rho*, computed between the rank orders

of the uncorrected *r*'s and the *r*'s corrected for attenuation is equal to .97, indicating that the unreliability of the *FinP* ratings had little effect on the rank order of the validities with which different variables were rated after the Intensive interview. However, differences in the reliability with which the variables were rated by the staff team during the Final Pooling Conference may be related to differences in the validity with which the variables were rated by the Intensive Interviewer. The *rho's* between the rank orders of the *FinP* reliability coefficients and the rank orders of the *Intens-FinP* correlations are equal to .69 for Scale B and .19 for Scale C, suggesting that for Scale B, trait differences in Intensive Interview validities would be smaller were the *FinP* ratings of equal reliability.

The *Intens-FinP r*'s were converted into *z*'s and tested for significance of differences between them in the manner described for the *Init-FinP r's*. Thirty-eight pairs of differences of the total of 189 pairs were found to be significant at the .05 level, indicating that differences do exist in the correlations between the *Intens* and *FinP* ratings.

3. *The contribution of the interview to the validity of ratings made after the Intensive interview.* The relative *Net Gain* figures for the Intensive interview (obtained by the method described for the Initial interview—Section II, A, 3 above) are shown in Column 9 of Table 1. The variables (B25, B30, C35, C38, C40, and C42) to which the Intensive interview makes the greatest relative contribution seem to be those concerned with *social relations with others in the professional work situation.* The variables to which the Intensive interview contributes very little (B23, B27, B29, C32, and C36) appear to be variables concerned with *effective intelligence* or *intellectual efficiency.* B23 (Social Adjustment) does not fall into this class, but this variable is relatively broad in definition and has many aspects which render it poorly ratable on the basis of a face-to-face conversation. Also, even relatively poorly adjusted individuals might be able to put up a façade for the interview period.

Role dominance may be a contributing factor to the apparent contribution of the Intensive interview. That is, those variables which show relatively high *Net Gains* may do so, not because of any contribution of the interview itself, but because the other two team members, believing the Intensive Interviewer to be the best informed on these variables, deferred to his judgments when determining the *FinP* ratings. In fact, five of the six variables with the highest relative *Intens Net Gains* were variables for which the Final Individual ratings (just before the Final Pooling Conference) made by the Intensive Interviewer correlated higher with the *FinP* ratings than did the Final Individual ratings made by either of the other two staff members (18). Of the five variables with the lowest relative *Intens Net Gains* only two were variables for which the Final Individual ratings of the Intensive Interviewer had higher *r's* with the *FinP* ratings.

C. Comparison of the Intensive and Initial Interviews

1. *Agreement between ratings made after the interviews.* In Column 10 of Table 1 are listed the correlations between the *Init* ratings and the *Intens* ratings. Thirteen of the 20 *r*'s were signifi-

cant at the .01 level. Two facts seem of especial interest here. First, the median *r* for Scale B is .19, while that for Scale C is .31. Apparently, the two interviewers agreed better with each other on ratings of performance five years in the future than they did on ratings of present personality traits. Second, for no variable did the ratings made after the interviews correlate with each other to a greater extent than either set of ratings correlated with the *FinP* ratings. In addition, it may be noted that the variables (B31, C32, and C36) on which the ratings made after the interviews showed the greatest agreement are among those which were most validly rated before the interview by both the Initial and the Intensive Interviewers and which gained least in terms of relative *net gains* from either of the interviews. That is, those variables to the rating of which the interviews contributed relatively little are the very ones on which there is most agreement after the interviews.

2. *Comparison of the Interviewers' ratings with respect to correlation with the FinP ratings.* In Columns 1 through 4 of Table 4 appear four sets of correlations between Interviewers' ratings and *FinP* ratings:

The *PreInit-FinP r*'s (Column 1) are the correlations between the *FinP* ratings and ratings (based on Credentials material only) made by the Initial Interviewer before the interview.

The *Init-FinP r's* (Column 2) are correlations between the *FinP* ratings and ratings by the Initial Interviewer after the one-hour Initial interview.

The *PreIntens-FinP r*'s (Column 3) are correlations between the *FinP* ratings and ratings by the Intensive Interviewer just prior to the interview. In making these ratings, the Intensive Interviewer had available Credentials material, Objective test data, Autobiographical material, and Projective protocols.

The *Intens-FinP r*'s (Column 4) are correlations between the *FinP* ratings and ratings by the Intensive Interviewer after a two-hour Intensive interview.

These four sets of correlations permit of six comparisons, each of which yields some interesting information.

a. The comparative validities of preinterview ratings based on different amounts of written material. When the *PreInit-FinP* correlations are compared with the *PreIntens-FinP,* it is evident that the latter are appreciably higher for all variables. When it is considered that the *PreInit* ratings were based on Credentials material while the *PreIntens* ratings were based on Credentials material plus considerable other psychological data, it seems indicated that the validity of personality-trait ratings will vary with the amount of psychological data made available to the raters.

b. The comparative validities of ratings made after the Initial and Intensive interview situations. The *Intens-FinP* correlations were consistently higher than the *Init-FinP* correlations. Whether the relatively greater validity of ratings after the Intensive interview situation can be attributed to the greater amount of psychological data available to the Intensive Interviewer, or to the greater length and depth of the Intensive interview, or to both factors, is a matter of speculation. The evidence presented below (Section II, C, 3) when the interviews are compared with respect to relative *net gains* suggests that most, if not all, of the apparent superiority of the Intensive interview situation is a result of the greater amount of written data available before the interview.

c. The comparative validities of ratings made before and after the interviews. Four comparisons may be made here. Two of these, the comparison of the validities of ratings before and after the Initial interview and before and after the Intensive interview have been made indirectly in the sections (II, A, 3, and II, B, 3, above) wherein the material concerning the *net gains* in validity attributable to the interview was presented, so it need be only stated here that for all variables, ratings made after each type of interview correlated higher with the *FinP* ratings than did ratings made before that interview. The third comparison, that between *PreInit-FinP* correlations and *Intens-FinP correlations,* merits little consideration since it merely substantiates the rather obvious hypothesis that ratings based on a thorough interview plus a comprehensive mass of written psychological data are more valid than ratings based on only a little written data without benefit of an interview.

The last comparison, that between *Init-FinP* correlations and *PreIntens FinP,* seems of such importance that it is worthy of consideration separately.

d. Comparison of ratings made after the Initial interview with ratings made before the Intensive interview. Proponents of the appraisal interview frequently state that the value of interviews lies in their flexibility. The argument goes that in the interview situation it is possible to follow up leads; to confirm or reject hypotheses regarding the personality dynamics of the interviewee; and thus to form much more accurate judgments than are possible on the basis of test data alone. Test scores, on the other hand, are rigid and inflexible. Even projective devices offer only hints to possible dynamics. Thus, test material may have some value in assessing the more superficial surface traits, but only by talking to the subject, by asking him questions, and by listening to his answers, may the underlying, more fundamental, source traits be accurately appraised.

Comparison of the *Init-FinP* correlations (Column 2 of Table 1) with the *PreIntens-FinP* correlations provides at least a partial test of this belief. The *Init* ratings were based on a one-hour interview (plus the background material available in the Credentials file). The *PreIntens* ratings were based entirely on written data (with no opportunity to talk to, or even see, the subject). The Scale B traits are source traits and hence supposedly more validly rated on the basis of an interview.

This belief is not upheld by the present study. The median *Init-FinP* correlation for Scale B traits is .42; the median *PreIntens-FinP* correlation is .49. The *Init* ratings are more valid for two of the Scale B traits; the *PreIntens* ratings are more valid for seven.

Approximately the same relationship existed for the Scale C variables. The median *Init-FinP r* is .46; the median *PreIntens-FinP r* is .57. The *PreIntens* ratings are more accurate for nine of the eleven variables.

Thus, *ratings based on only test and biographical data are more valid than ratings based on an interview plus the information contained in a credentials folder.*

3. *Comparison of the interviews with respect to relative net gains.* When the relative *net gains* in correlations between ratings which can be attributed to the interviews (Columns 8 and 9 of Table 1) are compared, the Intensive Interviewer seems to gain slightly more from the interview than does the Initial Inter-

viewer, although the difference is neither statistically nor practically significant. The median *net gain* is .24 for the Intensive Interviewer and .16 for the Initial Interviewer. The fact that the Intensive Interviewer appears to have been considered by the other two staff members as being in a somewhat better position to rate many of the variables renders even this small difference questionable.

III. DISCUSSION, CONCLUSIONS, AND SUGGESTIONS FOR FURTHER RESEARCH

THE PRESENT investigation indicates that interviews apparently do contribute to the validity with which personality variables are rated (i.e., the assessment raters typically made more valid ratings on the basis of an interview plus psychological data than they did on the basis of psychological data alone). That contribution is slight, however, and may be even less than has been demonstrated here. It should be remembered that two of the three criterion team members were the interviewers whose ratings were studied. Were the criterion measures truly independent—had none of the criterion team members been interviewers—it does not seem likely that the validity of the interviewers' ratings would have been as high as was found in this study. On the other hand, it is possible that had the interviews occurred first—with psychological data presented to the interviewer after he had already rated on the basis of the interview—their validity coefficients might have been higher, with the psychological test data yielding only a slight additional validity to the interview ratings already made. That is, the psychological data when presented first as in this investigation may have given the interviewers a set so that they were unable to change their ratings after the interview. It seems unlikely that any such persistence of initial impressions actually did decrease the effective validity of the ratings after the interviews for these reasons: (1) The Initial Interviewers, who had a minimum of psychological information (only the credentials files) available before the interview, did not gain more from the interview than did the Intensive Interviewers who had much psychological data available before the interview and thus had opportunity to form stronger initial impressions. (2) Ratings based on unstructured appraisal interviews, unless used in conjunction with some sort of objective psychological data, have been shown to be unreliable, and hence invalid. Additional research is necessary for a definite answer to this question. It is unfortunate that the assessment program could not have been designed to answer this and the similar questions arising when the validity of any of the many assessment techniques is investigated, but, of course, all possible permutations of the various techniques would require several hundred summers of assessments. Since but one assessment was possible it was actually designed on the basis of increasing cost of technique. Thus, credentials costing only postage were made available to the staff first, objective tests second, the interviews (fairly expensive since they required both students and interviewers to be present) later, and last the situation tests requiring groups of students and groups of staff members.

When the two types of interview situations are compared with respect to validity of ratings made by the interviewers after the interviews, it is apparent that the Intensive Interviewer was typically able to make ratings which were more valid than were ratings made by the Initial Interviewer. When the two types of interview situations are compared with respect to relative *net gains* this relationship still exists, but the difference becomes statistically insignificant. In addition, when it is considered that the Intensive Interviewer may have been con-

sidered by the other two team members to be the person in the best position to rate, a large portion of this difference in validity would seem to disappear.

There appears, then, to be little difference in the true relative contributions of the two types of interview situations. The Intensive interview, which was two hours in length and in which an attempt was made to probe for underlying personality dynamics, contributed little, if any, more than did the Initial interview, which lasted only one hour and which was ostensibly devoted to the gathering of rather superficial information.

Since differences in the interviews themselves cannot account for the differences in validity of ratings after the two interview situations, these differences would seem to be functions of the differences in the amount of psychological test data available to the interviewers for study before the interview. This is further evidenced by the fact that the *PreIntens* validities were higher than the *PreInit* validities.

Of equal significance is the fact that ratings based on the one-hour interview plus credentials (the *Init* ratings) correlated less well with the criterion ratings than did the *PreIntens* ratings—ratings which were based on a varied and comprehensive set of psychological data but without any interview. Apparently, skilled clinicians can use fairly complete psychological test data in the rating of personality traits slightly more effectively then they can use an interview, even when the interview is aided by historical material in the form of credentials. Unfortunately, neither set of ratings has much effective validity. The median validity of the *Init* ratings for all twenty variables is .45 and the median validity of the *PreIntens* ratings is .49. Such correlation coefficients account for less than 25% of the variance in the criterion ratings, and indicate the predictor ratings to be only about 25% more valid than chance ratings—this despite the fact that the criterion ratings are based on judgments of the same persons who interpreted the psychological data and conducted the interviews. If the Initial interview situation is in fact (as suggested in Section I, E) comparable to the typical personnel selection or college admission interview, it would appear these interviews have little actual value.

Combination of comprehensive psychological data and a two-hour probing interview results in Intensive interview ratings whose median correlation with the criterion ratings is about .63, accounting for about 40% of the criterion variance. This is a sizeable increase over the 20% accounted for by ratings based on psychological data alone, but still leaves the greater percentage of the criterion rating variance unaccounted for—a percentage which would probably be still greater were there not a personnel overlap between the persons making the predictor and the criterion ratings.

Validity coefficients of .65 are considered fairly satisfactory for group tests used for selection purposes, but for individual personality assessment somewhat higher validities are needed. Especially is this true in the areas of clinical psychology and psychiatry where decisions of basic importance to clients and patients must be made—and frequently are made on the basis of much less complete psychological data than were available in the present instance.

A. Conclusions

The conclusions to be drawn from this investigation have been implicit through-

out this and the preceding chapter. Explicitly stated they are:

1. The contribution of the unstructured appraisal interview to the validity of personality-trait ratings based on an interview plus psychological data is slight.

2. Longer interviews with the objective of uncovering personality dynamics contribute little, if any, more to the validity of personality-trait ratings than do shorter interviews whose main objective is the eliciting of information.

3. The more comprehensive the psychological data available, the more valid will be personality-trait ratings based on that data.

4. Personality-trait ratings based on fairly comprehensive psychological test data are more valid than are ratings based on interviews with only a little psychological data in the form of credentials material available.

5. At the present time, even skilled clinicians, having available for study and integration a wide variety of objective test data, projective protocols, credential material, and autobiographical data, plus data accruing from a probing face-to-face interview, do not appear able to make personality assessments, in the form of ratings of personality traits, which have sufficient validity to be as useful in individual case work as is desirable.

B. Suggestions for Further Research

This investigation, although it represents a definite contribution to the study of the validity of personality-trait ratings based on interviews, in that it provides an estimate of the increment in validity which may be attributed to the interview itself, has left at least three important questions unanswered:

1. The previously mentioned problem of whether the presentation of psychological data before the interview formed a "set" in the interviewer which prevented his post-interview ratings being changed sufficiently to reflect the true validity of the interview.

2. The interview as a selection instrument could not be evaluated here since the Final Pooled ratings of Scale C variables did not appear to be acceptable as selection criteria. A similar study is necessary to determine the incremental validity of interview ratings of job-success variables when those ratings are evaluated against acceptable criteria of job-success.

3. Although this investigation was able in some measure to estimate the incremental validity of the interview when used to assess personality variables, it is probable that such incremental validity has been overestimated inasmuch as the criterion measures themselves were determined by the persons who originally made the interview ratings. It would seem desirable, then, that another study be carried out so that the incremental validity of the interview in the assessment of personality-traits might be evaluated against truly independent criteria.

IV. SUMMARY

A. Objectives of This Study

The main objectives of this study were:

1. To determine the relative validity of ratings of personality traits based on two types of interview situations, differing in amounts and kinds of material available to the interviewer and in the length of the interview.

2. To determine the incremental validity of each of the two types of interview situations, i.e., how much more valid were interview ratings than other ratings made without benefit of the interviews?

Other related problems were also investigated.

B. Methods and Procedures

A total of 128 male first-year graduate students majoring in clinical psychology at some 30 universities were interviewed in two types of interview situations; one, an Initial Interview lasting one hour, with only credentials material available before the interview for study by the interviewer; and two, an Intensive Interview lasting two hours, with very comprehensive psychological test data available for study by the interviewer before the interview. The interviews were conducted by 16 clinicians (2 psychiatrists and 14 psychologists) who had had considerable prior interviewing experience, each interviewer acting interchangeably as Initial Interviewer and as Intensive Interviewer.

The subjects were rated by the interviewers before each interview on nine personality-trait variables and eleven "future performance" variables, and after the interview on 31 personality-trait variables and eleven "future performance" variables. Four sets of ratings were thus available for each subject: the Pre-Initial interview ratings, based on credentials material alone; the Initial interview ratings, based on credentials material plus an Initial interview; the Pre-Intensive interview ratings, based on comprehensive written psychological data; and the Intensive interview ratings, based on comprehensive psychological data plus an Intensive interview.

The validity of each of the four sets of ratings (for each variable) was estimated by correlating each set of ratings with criterion ratings arrived at by a team of three psychologists who had intensively studied each subject for a period of one week. The incremental validity of each type of interview was estimated by comparing the validity of ratings made just before and just after the interview, and, in addition, by comparing ratings based on the Initial interview situation with the Pre-Intensive ratings based on comprehensive written psychological data but without an interview. These estimates of validity and incremental validity are probably spuriously large, due to personnel overlap between interviewers and criterion teams.

C. Chief Findings

In summary it was found that:

1. Ratings made after both types of interview correlated significantly with the criterion ratings.

2. Significant differences existed in the relative validity with which the different variables were rated.

3. Criterion unreliability decreased only slightly the validity of interview ratings.

4. Ratings made after the Intensive interview were more valid than ratings made after the Initial interview.

5. Ratings made after each type of interview correlated significantly with ratings made after the other type of interview, but these correlations were lower than correlations between each set of ratings and the criterion ratings.

6. Ratings after each type of interview correlated higher with the criterion measures than did ratings made before the interview, but there appeared to be no differences between the two interview situations (when role dominance is considered) in terms of gain in validity of ratings after the interview over ratings before the interview.

7. Ratings made by the Intensive Interviewer before the interview (based on written material alone) were slightly more valid than ratings made by the Initial Interviewer after the interview.

D. Conclusions

In summary it was concluded that:

1. The incremental validity of the unstructured interview when used in the assessment of personality is slight, and is even slightly negative when ratings based on an interview plus a minimal amount of psychological data are compared with ratings based on comprehensive written data but without an interview.

2. Longer, probing interviews have little more incremental validity than short "information-eliciting" interviews.

3. The more comprehensive the psychological data available, the more valid will be personality-trait ratings based on such data.

4. At the present time even skilled clinicians, basing their judgments on comprehensive psychological data plus an interview, do not appear able to make personality-trait ratings with sufficient validity to be as useful in individual case work as is desirable.

APPENDIX

RATING SCALE DEFINITIONS

SCALE B (Variables Nos. 23-31)

Note: For ratings on this scale, 1 = left side or low, 8 = right side or high.

Since many of the following attributes (#23-30) are broad factors, it is unlikely that any person will fit all the phrases grouped together at one pole of a given variable. Note also that for some items neither extreme necessarily represents a desirable attribute.

23. *Social Adjustment:* How well does he adjust to varied interpersonal situations? (Includes sexual adjustment only as it affects social adjustment in general.)

Acts without consideration for feelings of others; often rejected by others, often appears aloof, hostile, or irritable.	Actively considers feelings of others; readily gains acceptance in interpersonal relationships; maintains a friendly and likeable manner.

24. *Appropriateness of Emotional Expression:* How appropriate are his emotional responses to the situation?

Fails to adapt his emotional responses to the needs of the situation; shows disorganized or overly constricted emotional responses.	Shows emotional responses of a quality and intensity befitting the situation; reacts spontaneously but appropriately; shows well-integrated and flexible patterns of emotional behavior.

25. *Characteristic Intensity of Inner Emotional Tension:* How intense is his inner emotional life as *inferred* from all available clues?

Inner emotional life characterized by a minimum of persistent internal tensions.	Has strongly repressed emotional drives resulting in inner turmoil; great inner conflict and strong pent-up emotions.

26. *Sexual Adjustment:* To what degree do his sexual needs and activities affect his overall adjustment?

His sexual needs and activities seriously interfere with his overall adjustment.	His sexual needs and activities definitely enhance his overall adjustment.

27. *Motivation for Professional Status:* How strong is his drive for the status-rewards of a professional career?

28. *Motivation for Scientific Understanding of People:* How strong are his drives toward acquiring the facts, theories, and skills necessary for the scientific understanding of individual human beings?

29. *Insight into Others:* How much insight does he have into the attitudes, emotions, and motivations of others?

Interprets behavior at its face value; insensitive to any but gross differences in behavior; does not develop any integrated understanding of behavior or of people.	Has good awareness of underlying dynamics of behavior; is sensitive to subtle nuances of behavioral responses; is able to develop integrated understanding of the behavior of people.

30. *Insight into Himself:* How much insight does he have into the underlying dynamics of his own attitudes, emotions and motivations?

31. *Quality of Intellectual Accomplishments:* What is the characteristic quality of his intellectual output?

Intellectual work is characteristically of low quality.	Characteristically produces intellectual work of high quality.

SCALE C—CRITERION SKILLS (Variables Nos. 32-42)

Note: For ratings on this scale, 1 = low, 8 = high.

Ratings on No. 32 refer to performance in graduate school; ratings on Nos. 33-42 refer to student's performance five years hence (i.e., after one year of experience past the Ph.D.).

What will be his level of competence or skill in the varied aspects of:

32. *Academic Performance* (during next three or four years): How well will he:

Effectively master course content, successfully complete courses in general psychology, clinical psychology, statistics, and related fields; satisfy language requirements for the doctorate; pass general examinations.

33. *Clinical Diagnosis:* How well will he:

Recognize dynamics underlying particular responses in both objective and projective tests, observe significant interrelationships among responses, relate findings to case history and other clinical data.

Elicit from the patient information required for mental status examinations and case histories; ascertain and evaluate attitudes and incidents of psychological significance in the patient. Synthesize clinical findings to arrive at an integrated picture of personality development, structure, and function.

34. *Individual Psychotherapy:* How effectively will he: Conduct various types of individual psychotherapy.

35. *Group Psychotherapy:* How effectively will he: Conduct various types of group psychotherapy.

36. *Research:* How well will he:

Recognize and define important research problems in clinical psychology; critically evaluate and apply the research findings of others; think with originality and scientific rigor; employ appropriate experimental design and statistical methods; grasp practical implications of findings; present results and conclusions in clear, comprehensive, and well-organized form.

37. *Administration:* How well will he:

Plan and develop psychological programs; make proper administrative decisions; delegate responsibility appropriately; elicit cooperation from subordinates and superiors; maintain high morale among his staff; carry out or direct an appropriate public relations program.

38. *Supervising Clinical Psychologists:* How well will he:

Carry out the professional supervision of subordinates assigned to him for duty and on-the-job instruction; assign their duties; evaluate their performance; instruct them in clinical techniques; perform other aspects of in-service training.

39. *Teaching Psychology* (in a College or University): How well will he:

Teach college courses in general psychology; motivate students; present concepts and procedures; stimulate critical thinking about and integration of course materials; evaluate the products of learning.

40. *Professional Interpersonal Relations:* How well will he:

Work cooperatively with superiors, subordinates, members of the mental team, and other professional personnel concerned with the patient's welfare; participate in the give-and-take of staff conferences; contribute to group decisions.

41. *Integrity of Personal and Professional Behavior:* How well will he:

Recognize and fulfill professional responsibilities; live up to personal commitments; show loyalty to professional obligations in the event of outside pressure or promise of personal gain; maintain discretion concerning professional matters; appropriately conform with commonly accepted standards of moral and social behavior; refrain from coloring facts, evasion, lying, etc.

42. *Overall Suitability for Clinical Psychology:* In view of his assets and liabilities, how well will he be able to:

Carry out the several duties—diagnosis, therapy, and research—specified for the position of clinical psychologist (P-4 and above) in the Veterans Administration.

BIBLIOGRAPHY

1. *Assessment of men.* OSS Assessment Staff. New York: Rinehart, 1948.
2. BINGHAM, M. V. D., and MOORE, B. V. *How to interview.* New York: Harper, 1941.
3. CATTELL, R. B. *Description and measurement of personality.* Yonkers, N.Y.: World Book Co., 1946.
4. CLARK, E. B. Value of student interviews. *J. Person. Res.,* 1926, **5,** 204-207.
5. ESCALONA, SYBILLE. *Progress report: research project on the selection of medical men for psychiatric training.* Topeka: Menninger Foundation, 1948. (Mimeographed and privately circulated.)
6. FEARING, FRANKLIN. The appraisal interview: a critical consideration of its theory and practice with particular reference to the selection of public personnel. In *Studies in personality.* New York: McGraw-Hill, 1942.
7. FEARING, F., and FEARING, F. M. Factors in the appraisal interview considered with particular reference to the selection of public personnel. *J. Psychol.,* 1942, **14,** 131-153.
8. FISKE, DONALD W. Consistency of the factorial structures of personality ratings from different sources. *J. abnorm. soc. Psychol.,* 1949, **44,** 329-344.
9. HOLLINGWORTH, H. L. *Vocational psychology and character analysis.* New York: Appleton, 1929.
10. HOVLAND, C. I., and WONDERLIC, E. F. Prediction of industrial success from a standardized interview, *J. appl. Psychol.,* 1939, **26,** 537-546.
11. KELLY, E. L., and FISKE, D. W. *The selection of clinical psychologists: progress report and preliminary findings of the research project on the selection of clinical psychologists.* Ann Arbor, Michigan, 1948. (Lithographed and privately circulated.)
12. LINDQUIST, E F. *Statistical analysis in educational research.* Boston: Houghton Mifflin, 1940.
13. MOSS, F. A. Scholastic aptitude tests for medical students, *J. Ass. Amer. Med. Colleges,* 1931, **6,** 1-16.
14. NEWMAN, S. H., BOBBITT, J. M., and CAMERON, D. C. The reliability of the interview method in an officer candidate evaluation program. *Amer. Psychologist,* 1946. **1,** 103-110.
15. SCOTT, W. D., BINGHAM, M. V., and WHIPPLE, G. M. The scientific selection of salesmen. *Salesmanship,* 1916, **4,** 106-108.
16. SOSKIN, W. F. Standardized situations as a means of personality appraisal. *Amer. Psychologist,* 1948, **3,** 271.
17. SYMONDS, P. M. *Diagnosing personality and conduct.* New York: Appleton-Century, 1931.
18. TUPES, E. C. *The effect of role dominance on estimates of validity of interviewers' ratings* (in preparation).
19. TUPES, E. C. *Changes in means and standard deviations of personality-trait ratings after interviews* (in preparation).

Psychological Monographs

General and Applied

An Evaluation of Personality-Trait Ratings Obtained by Unstructured Assessment Interviews

By

Ernest C. Tupes

Edited by Herbert S. Conrad
Published by The American Psychological Association

Psychological Monographs:

General and Applied

MANUSCRIPTS should be sent to the Editor. For suggestions and directions regarding the preparation of manuscripts, consult the following article: CONRAD, H. S. Preparation of manuscripts for publication as monographs. *J. Psychol.*, 1948, 26, 447-459.

Because of lack of space, the *Psychological Monographs* can print only the original or advanced contribution of the author. *Background and bibliographic materials must, in general, be totally excluded,* or kept to an irreducible minimum. Statistical tables should be used to present only the most important of the statistical data or evidence.

CORRESPONDENCE CONCERNING BUSINESS MATTERS (such as subscriptions and sales, change of address, author's fees, etc.) should be addressed to: DR. DAEL WOLFLE, American Psychological Association, 1515 Massachusetts Ave., N.W., Washington 5, D.C.

PSYCHOLOGICAL MONOGRAPH

#270

1945

ON PROBLEM SOLVING

BY

KARL DUNCKER

$2.50

This popular monograph is #270 of the Psychological Monograph series. It has been reprinted so that it is again available.

American Psychological Association

1515 Massachusetts Avenue N.W.

Washington 5, D. C.